The effect of missing data on classification quality

Feldman, Michael ; Even, Adir ; Parmet, Yisrael

Abstract: The field of data quality management has long recognized the negative impact of data quality defects on decision quality. In many decision scenarios, this negative impact can be largely attributed to the mediating role played by decision-support models - with defected data, the estimation of such a model becomes less reliable and, as a result, the likelihood of flawed decisions increases. Drawing on that argument, this study presents a methodology for assessing the impact of quality defects on the likelihood of flawed decisions. The methodology is first presented at a high level, and then extended for analyzing the impact of missing values on binary Linear Discriminant Analysis (LDA) classifiers. To conclude, we discuss possible directions for extensions and future directions.

THE EFFECT OF MISSING DATA ON CLASSIFICATION QUALITY

(Research in Progress)

Michael Feldman
Ben-Gurion University of the Negev, Israel
fmichael@bgu.ac.il

Adir Even
Ben-Gurion University of the Negev, Israel
adireven@bgu.ac.il

Yisrael Parmet
Ben-Gurion University of the Negev, Israel
iparmet@bgu.ac.il

Abstract: The field of data quality management has long recognized the negative impact of data quality defects on decision quality. In many decision scenarios, this negative impact can be largely attributed to the mediating role played by decision-support models - with defected data, the estimation of such a model becomes less reliable and, as a result, the likelihood of flawed decisions increases. Drawing on that argument, this study presents a methodology for assessing the impact of quality defects on the likelihood of flawed decisions. The methodology is first presented at a high level, and then extended for analyzing the impact of missing values on binary Linear Discriminant Analysis (LDA) classifiers. To conclude, we discuss possible directions for extensions and future directions.

Key Words: Data Quality, Missing Values, Decision Making, Classification, Linear Discriminant Analysis

INTRODUCTION AND BACKGROUND

The common saying "Garbage in Garbage Out" reflects a key concern in the field of data quality management (DQM) – the negative impact of data quality (DQ) defects on decision making (Redman, 1996; Shankaranarayanan and Cai, 2006; Liu et al., 2010). This study explores that impact through the mediating role played by decision-support models, arguing that a wrong decisions are often the result of an unreliable model that was a built from low-quality data. Decision-making is often supported by a model (Shim et al., 2002) - a form of representation (e.g., theoretical, analytical, visual, statistical) that describes phenomena or behaviors in the real-world. Such a model permit prediction of future behavior to an extent and, by that, assists with the formation of decisions and actions. Following this notion, Decision-Support Systems (DSS) provide the infrastructure and the utilities for building, applying and evaluating models that aid the decision-maker.

Recent years have witnessed a major transition toward decision-making culture that is based on data collection and analysis (Davenport, 2006). This transition can be associated with the growing popularity of Business Intelligence and Data Warehousing (BI/DW) systems – DSS that rely on the collection and integrating data from diverse resources (Davenport, 2006). Data repositories, in BI/DW systems and others, are often subject to DQ defects – such as missing, inconsistent, and/or inaccurate data values. Such defects might create a biased view of the real-world and, consequently, lead to flawed decisions and actions. A plethora of studies (e.g., Redman, 1996; Heinrich et al., 2009; Even et al., 2010) have described real-world scenarios in which defected data led to wrong decisions and major damages. The goal of this study is to contribute some insights into the mechanisms that may further explain that link.

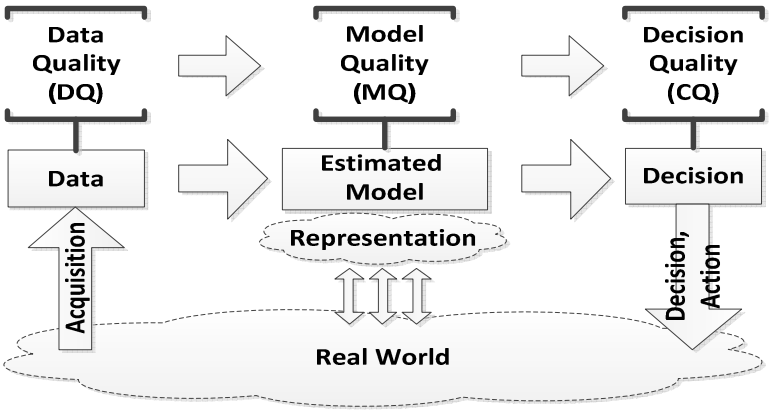

Figure 1: A Decision Process

Our methodology is conceptualized along three key stages of a typical data-driven decision process (Shim et al., 2002), and the associated quality assessments (Figure 1):

- **Data Quality (DQ):** Organizational data resources are built through ongoing complex processes of data acquisition, transfer and storage, during which they might become subject to DQ defects (Ballou et al., 1998; Parssian et al., 2004). Those data resources can support a variety of usages (Davenport, 2006, Even and Shankaranarayanan, 2007) – in this study we particularly observe the use of data for constructing and estimating models for decision-making support. DQ can be assessed along multiple dimensions, each reflecting a different type of data quality defects (Pipino et al., 2002, Even and Shankaranarayanan, 2007) – e.g., currency that reflects data that is not up-to-date, and accuracy that reflects incorrect values. This study addresses the impact of missing values – a common type of data quality defects, which is typically associated with the DQ dimension of completeness (Even et al., 2010). Data values may be missing due to reasons such as poorly designed data-entry screens, details that were not available (or not provided on purpose) at the time of data collection, database storage and update failures, or processing errors (Redman, 1996). This study focuses on missing completely at random (MCAR) patterns (Little, 1987), where missing data in one attribute does not depend on missing-value behavior in other attributes. Other missing-value patterns, such as missing at random (MAR) and not missing at random (NMAR), may assume some dependency between missing values. Such patterns should be further explored in future extensions to this study.
- **Model Quality (MQ):** The number of data items is often very large; hence, in many decision scenarios, data cannot be used as is. It is more common to use the data for constructing models that reflects real-world behavior in more compact and aggregated forms (e.g., formulas, charts, reports, digital dashboards, and the subject of this study – statistical classification models) that let a decision maker understand and analyze certain phenomena and behaviors. Model complexity and reliability may significantly affect decision making (Shim et al., 2002, Blake and Mangiameli, 2011). We interpret MQ is an assessment of model goodness – the extent to which our model reflects the true reality in a reliable manner. It is likely that with a higher rate of data quality defects (reduced DQ), the estimated model will provide a less reliable representation of reality (reduced MQ).
- **Decision Quality (CQ):** Models can serve as an input to decision-makers for gaining insights on how the real-world behaves, making some assessments and predictions, and act accordingly. The link between data quality and decision correctness, which has been explored in a variety of studies (e.g., Askira-Gelman, 2011, Blake and Mangiameli, 2011), is often complex and difficult to assess. We define CQ as the extent to which the decisions are correct. It is reasonable to assume that a flawed model might lead to misconceptions, flawed insights and hence wrong decisions – what motivates our claim that CQ is affected by MQ; hence, also by DQ.

In this study, we focus on classification – decision scenarios in which we associate a certain object, behavior, or situation with one category (or class) among a set of choices. Many decision scenarios, in different contexts, can be interpreted as classifications – e.g., replenishing inventory items (Davenport, 2006), assigning a customer to a segment (Even et al., 2010), or medical decisions, based on patient diagnostics (Session and Valtorta, 2009). Misclassification might damage reputation (e.g., misclassifying customers as "unimportant"), result in losses (e.g., investing in "overestimated" assets), or even threaten life (e.g., failing to detect hazardous medical conditions). Classifications often rely on models that can help associating a certain object with a certain class among a given set of choices – e.g., Distance-Based classifiers, k-Nearest-Neighbors (kNN), and Bayesian Classifiers (Duda and Hart, 2001). Classification models are often estimated (or "trained") from a dataset. If the "training" dataset suffers from DQ defects – the estimated classifier is likely to be biased; hence, with a higher likelihood, the resulting decisions will be flawed. In this study we chose to evaluate our methodology with a relatively simple but common classifier – the binary Linear Discriminant Analysis (McLachlan, 1992). The next section introduces a methodology that links the quality levels described above – data, model, and decision - and highlights the relationships among them in the context of classifiers. The methodology is further developed for binary LDA – but some of the evaluation and measurement methods applied can be used in broader contexts. The concluding section summarizes the key contributions of our study, highlights its limitations, and proposes possible extensions and directions for future research.

THE IMPACT OF INCOMPLETENESS ON CLASSIFIERS

This section develops a methodology for assessing the impact of data quality (DQ) on model and decision quality (MQ and CQ, respectively). The methodology (Figure 2) consists of the following components:

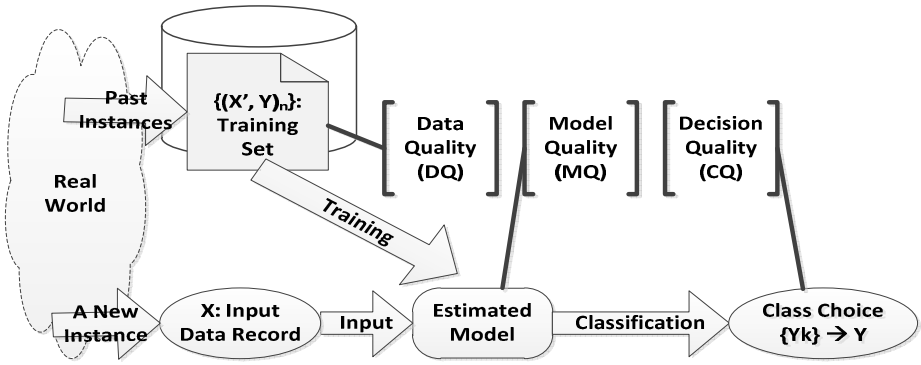

Figure 2: The General Methodology

Training Sets and Data Quality Measurement (Q^D): the data stored in organizational repositories can be used for the estimation of classification models. Following common terminology (Duda and Hart, 2001), we refer to the process of estimating the model "training" and to the dataset $\{(X, Y)_n\}$ used to estimate the model as a "training set". The annotation reflects N records (indexed 1..N), where X is a vector of M attributes (indexed 1.. M), each reflecting a certain property of a real-world instance. The Y component is a 1..K integer that associates the record with one among K classes. Following common DQ measurement schemas (Even and Shankaranarayanan, 2007), each record is associated with a Q_n measurement of completeness - 0, if one or more attribute values (or the entire record) are missing (i.e., NULL), 1 is the record is complete. The quality of the entire dataset Q^D, in terms of completeness, is defined as the rate of non-missing values, where Q^D=1 reflects a complete training set:

$$Q^D = \frac{1}{N}\sum_{n=1}^{N} Q_n, \quad 0 \le Q^D \le 1 \qquad (1)$$

Classification Models, and Model Quality Measurement (Q^M): A classifier can be described, in general, as a function M(X)=Y that maps an M-dimensional input vector X, which reflects a real-world instance to be classified, to an output integer Y=1..K associated with a class within a K-class set. In the decision scenarios that we discuss, the classifier parameters have to be estimated from a training set, as discussed above. With an "infinite" number of random sample (i.e., a very large N), the estimates of model parameters are likely to be accurate and reliable. However, with a smaller number of samples, the likelihood of misestimating parameters is higher and so is the likelihood of classification errors.

The confidence interval (CI) is a common approach for assessing the reliability of estimated model parameters. For example, when estimating a certain parameter A from a training set – the estimated value â is not necessary the true one. CI assessment would allow us to assume that "with a confidence of g% the true value of A resides within the CI of [â- Δ_1, â+ Δ_2]". Obviously – the smaller are the CI's for all parameters, the more reliable is the classification model. Further, with classification models that involve CI assessment, it can be shown that the CI gets smaller with a higher N. Adopting the CI-assessment concept - we take L, the length of the confidence interval as a measure for model quality (i.e., if the confidence interval is defined by [â-Δ_1, â+ Δ_2], then L = Δ_1+ Δ_2). The model-quality metric has to be defined for each model parameter A. It has to consider the desired target confidence level ρ, the number of samples N in the complete dataset, and the missing value rate (as reflected by Q^D):

$$Q_A{}^M\left(\rho, N, Q^D\right) = L_A\left(\rho, N * Q^D\right) \qquad (2)$$

Where

A -	The model parameter under evaluation
ρ -	The target confidence level
N -	The number of samples in the complete training dataset
Q^D -	The data quality level (i.e., the rate of non-missing values)
$L_A(x, y)$ -	The CI length for parameter A, given target confidence level y, and x samples

Confusion Matrix, and Decision Quality Measurement (Q^C): The classification output Y is an integer in the range of [1..K], which reflects an association to the input record (or vector) X to one class within a K-class set. A classification is said to be correct if an instance that belongs to class k is indeed classified to class k, and incorrect otherwise. With binary classifiers (i.e., K=2), in which the output is either positive (Y=1) or Negative (Y=0), it is common to assess classification performance with the 2-way confusion matrix (Table 1) – a Positive item that was classified as Positive is considered as "True Positive" (TP), and so on (Han and Kamber, 2006). The total number of instance per quadrant (N_{TP}, N_{FP}, N_{FN}, N_{TN}, respectively, where $N_{TP}+N_{FP}+N_{FN}+N_{TN} = N$), are commonly used for assessing the following classification quality metrics, and possibly others:

- **Classification Accuracy ($Q^{C/A}$),** reflecting the rate of items classified correctly: $(N_{TP} + N_{TN}) / N$
- **Classification Precision ($Q^{C/P}$),** reflecting correctness within positive results: $N_{TP} / (N_{TP} + N_{FP})$
- **Classification Sensitivity ($Q^{C/S}$),** reflecting the ability to detect positive results: $N_{TP} / (N_{TP} + N_{FN})$
- **Classification Specificity ($Q^{C/F}$),** reflecting the ability to detect negative results: $N_{TN} / (N_{TN} + N_{FP})$

Real-World Class	Classification	
	1	**0**
1	True Positive (TP)	False Negative (FN)
0	False Positive (FP)	True Negative (TN)

Table 1: Binary Classification Assessment with 2-Way Confusion Matrix

A more general formulation of classifier-performance assessment, which can also address classifications with a larger number of classes (K>2), uses a confusion matrix (Table 2). The a-priory probabilities $\{V_1 \ldots V_K\}$ reflect the real-world distributions of classed ($\sum_{k=1..K} V_k=1$). The matrix items $\{W_{i,j}\}$ ($(\sum_{j=1..K} W_{i,j}=1$) reflects the probability of a real-world instance that belongs to class i to be classified as class j (a correct classification if i=j, incorrect otherwise). Accordingly, the decision quality Q^C is defined as the overall likelihood of correct classification (similar to "classification accuracy" for the binary classification case):

$$Q^C = \sum_{k=1}^{K} V_k W_{k,k}, \quad 0 \le Q^C \le 1 \qquad (3)$$

Real-World Class	A-Priory Probability	Classification			
		1	**2**	**…**	**K**
1	V_1	$W_{1,1}$, $U_{1,1}$	$W_{1,2}$, $U_{1,2}$	…	$W_{1,K}$, $U_{1,K}$
2	V_2	$W_{2,1}$, $U_{2,1}$	$W_{2,2}$, $U_{2,2}$	…	$W_{2,K}$, $U_{2,K}$
…	…	…	…	…	…
K	V_K	$W_{K,1}$, $U_{K,1}$	$W_{K,2}$, $U_{K,2}$	…	$W_{K,K}$, $U_{K,K}$

Table 2: K-Class Confusion Matrix, Including Relative Costs

An enhanced definition of Q^C may take into account the relative classification value, assuming that certain classification errors are possibly more severe than others. The parameters $\{U_{i,j}\}$ in the weighted confusion matrix (Table 2) reflects that relative value of classifying an item that belongs to real-world class i as j. We assume that all the diagonal values are non-negative $U_{i,i} \ge 0$ (i.e., correct classification cannot cause a damage), and that that for each i and j, $U_{i,i} \ge U_{i,j}$. This means that misclassification cannot have a higher value than a correct classification (otherwise, we would have adjusted the classifier to "misclassify"). However, misclassification might have a negative value – i.e., a certain costly damage to the overall performance (i.e., Ui,j can be negative if i≠j). Following these assumptions, the decision quality Q^C definition can be adjusted to:

$$Q^C = \frac{\sum_{k=1}^{K} V_k \sum_{j=1}^{K} W_{k,j} U_{k,j}}{\sum_{k=1}^{K} V_k U_{k,k}} \le 1 \qquad (4)$$

Notably the denominator in that expression $U^{max} = \sum_{k=1..K} V_k U_{k,k}$ reflects the expected value from a single classification act, with no classification errors. Hence, Q^C reflects the ratio between the expected value with some misclassification and U^{max}. As the value of some misclassification can be negative, Q^C might turn out to be negative too (e.g., in case that some likelihood exists for very costly misclassification). When all the diagonal values are equal $U_{k,k} = U$, and when all other non-diagonal values are 0 (i.e., no value, and no damage), the Q^C expression in Equation 4 becomes identical to Equation 3.

A special treatment is needed for the case where the diagonal values are all 0, but some non-diagonal

values are negative - i.e., $U_{i,j}=0$ for i=j, $U_{i,j} \leq 0$ for i≠j. This case reflects a decision scenario in which there is no value associated with correct classification, but there is some damage associated with misclassification. In that case, instead of measuring decision quality as defined earlier, it would be more reasonable to measure the decision cost C^C:

$$C^C = \sum_{k=1}^{K} V_k \sum_{j=1}^{K} W_{k,j} U_{k,j} \leq 0 \qquad (5)$$

The decision quality and cost discussed so far may rely on the number of samples N in the training set. Even with an "infinite" number of samples (i.e., a very large N), the model may still have some classification errors due to possible overlaps between classes (as shown later for LDA classifiers). With a smaller, "finite" number of samples – the classifier's performance is likely to degrade further. We now define the decision quality $Q^C(N)$, as a function of the number of samples N. The upper limit Q^{C^*} reflects the best possible decision quality for a classifier that was estimated with an "infinitely large" number of sample and CI→0. Similarly, we define the decision cost $C^C(N)$ as f function of the sample size. The lower limit C^{C^*} reflects the lowermost decision cost for a given classifier, with very large N, and CI →0.

$$Q^{C^*} = Lim_{N\to\infty} Q^C(N), \qquad C^{C^*} = Lim_{N\to\infty} C^C(N) \qquad (6)$$

The metrics developed so far, within the measurement methodology introduced in this section, were defined in a general manner that permits their usage in many classification scenarios. However, we suggests that with further analytical development, such metrics can become even stronger tools for assessing and predicting DQ, MQ, and CQ behavior, and setting DQ policies accordingly. In the following section we demonstrate such an extension for the commonly-used, LDA classifiers.

DEVELOPMENT AND EVALUATION FOR BINARY LDA CLASSIFIERS

The binary Linear Discriminant Analysis (LDA) classifier (McLachlan, 1992; Duda and Hart, 2001) assigns an input vector X to either class Y_0 or class Y_1. For terminology convenience, and with no loss of generality, we term one class as "positive" and the other as "negative" and annotate them with "1" and "0" respectively. The LDA assumes that two classes reflect normally-distributed populations, with a different mean per class (μ_0 and μ_1 respectively), but with the same covariance matrix $\sum$. The LDA classifies a vector X (all attributes are continuous) to Y_0 or Y_1 by calculating a Cartesian product between X and a separation hyper-plane W and comparing the result to a threshold value *A*:

$$W \bullet X > A, \quad where \quad W = \Sigma^{-1}(\mu_1 - \mu_2) \qquad (7)$$

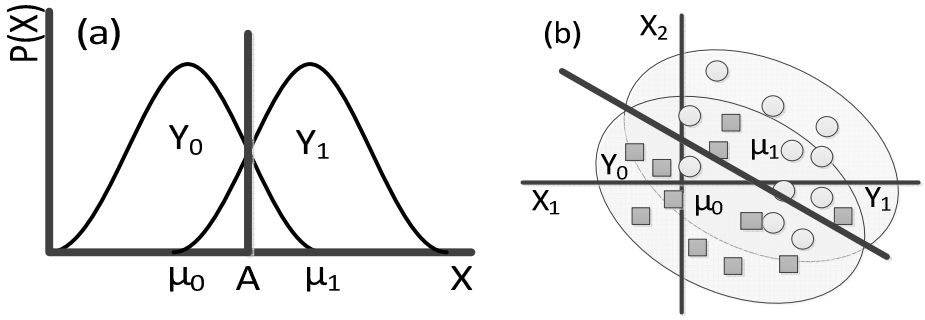

Figure 3: LDA Classifiers for (a) 1-dimensioal space, and (b) 2-dimensional space

Figure 3a shows a binary LDA classifier for a scalar ("1 dimensional") input, in which case the classification rule can be simplified to: X is classified as Y_1 if X > A, or classified as Y_0 otherwise (Again, with no loss of generality, we assume that the class with the higher mean is the "positive", while the class with the lower mean is "negative"). Figure 3b show a binary LDA classifier for a 2-dimensional input vector.

Both examples highlight the fact that the binary LDA is not a perfect classifier – some misclassifications may occur, as the populations of the two classes may overlap to an extent. However, it can be shown that given the parameters of the two distributions – the LDA classifier defines the optimal linear separation in terms of minimizing the likelihood of error. To demonstrate our evaluation concept, and highlighting the potential contribution, the rest of this section develops further the scalar (1-dimensionl) case. In the concluding section we will discuss a few extensions currently under research.

As summarized in Table 3, Y_1 ("positive") and Y_0 ("negative") are with a-priory probabilities of $V_1=V_0=0.5$. Each class reflects a Normally-distributed population with different means $\mu_1>\mu_0$ but the same STDEV σ. We consider a case where there's no positive value to correct classification, but some known cost U of misclassifications (The cost is identical for "False Positive" and "False Negative). With some probability W_{TP} a "positive" item can be classified correctly as "positive", and with some probability $W_{FN}=1-W_{TP}$ as "negative" ($W_{TP}+W_{FN}=1$). Similarly, with some probability W_{TN} a "negative" item can be classified correctly as "negative", and with some probability $W_{FP}=1-W_{TN}$ as "positive".

Class	**A-Priory Probability**	**Distribution Function**	**Classification**	
			1 – Positive	**0 – Negative**
1 - Positive	$V_1 = 0.5$	$P_1 \sim N(\mu_1, \sigma)$	**True Positive:** W_{TP}, 0	**False Negative:** W_{FN}, U
0 – Negative	$V_0 = 0.5$	$P_0 \sim N(\mu_0, \sigma)$	**False Positive:** W_{FP}, U	**True Negative:** W_{TN}, 0

Table 3: The Confusion Matrix, for the Binary LDA Case

The LDA model, in that case, has one parameter only – the threshold A that defines the classification rule (a new instance x, with unknown classification, is classified as "positive" if x>A, or "negative" otherwise). Based on the assumptions above, it can be shown that with known distribution parameters (μ_1, μ_0, and σ), the optimal threshold value, in terms of maximizing classification accuracy, is $A=0.5*(\mu_0+\mu_1)$, with a confidence interval of $CI_A=0$ (as the distribution parameters are known, and not estimated). The probabilities of correct classifications versus misclassification can be calculated accordingly as follows:

$$\begin{aligned} W_{TP} &= 1-\Phi\left((A-\mu_1)/\sigma\right) = 1-\Phi\left(\left(\frac{\mu_0+\mu_1}{2}-\mu_1\right)\Big/\sigma\right) \\ &= 1-\Phi\left((\mu_0-\mu_1)/2\sigma\right) = \Phi\left((\mu_1-\mu_0)/2\sigma\right) \end{aligned}$$

$$W_{FP} = 1-W_{TP} = 1-\Phi\left((\mu_1-\mu_0)/2\sigma\right) = \Phi\left((\mu_0-\mu_1)/2\sigma\right) \quad (8)$$

$$\begin{aligned} \textit{Due to symmetry} &: W_{TN} = W_{TP} = \Phi\left((\mu_1-\mu_0)/2\sigma\right), \\ & W_{FN} = W_{FP} = \Phi\left((\mu_0-\mu_1)/2\sigma\right) \end{aligned}$$

(Φ – Cumulative Normal Distribution)

The expected decision quality (Equation 3) for this case is:

$$Q^C = Q^{C^*} = V_1 * W_{TP} + V_0 * W_{TN} = \Phi\left((\mu_1-\mu_0)/2\sigma\right) \quad (9)$$

It can be shown that with known distribution parameters (μ_1, μ_0, and σ), the expression in equation 9 would be the best possible decision quality that can be obtained (hence, Q^{c^*}). With $\mu_1-\mu_0 \rightarrow 0$, and/or with $\sigma \rightarrow \infty$, $Q^{C^*} \rightarrow 0.5$ (a random "flip of a coin"). With $\mu_1 >> \mu_0$, and/or with $\sigma \rightarrow 0$, $Q^{C^*} \rightarrow 1$. The expected decision cost (Equation 5), in that case, would be:

$$C^C = C^{C^*} = U * \left(1-\Phi\left((\mu_1-\mu_0)/2\sigma\right)\right) = U * \Phi\left((\mu_1-\mu_0)/2\sigma\right) \quad (10)$$

Again, with known distribution parameters, this would be the lowest possible decision cost (hence, C^{c^*}). With $\mu_1 - \mu_0 \rightarrow 0$, and/or with very large σ, $C^{C^*} \rightarrow 0.5U$. With $\mu_1 >> \mu_0$, and/or with $\sigma \rightarrow 0$, $C^{C^*} \rightarrow 0$.

Parameter Estimation and Model Quality Metric for the Binary LDA Classifier

So far, the development reflected classifier parameters that are known in advance – however, in the decision scenarios that we discuss, the parameters μ_1, μ_0, and σ have to be estimated from a "training set" – $\hat{\mu}_1$, $\hat{\mu}_0$, and $\hat{\sigma}$, respectively. At full size, our "training set" has N samples for each class (a total of 2N). Some values are missing from that training set, hence a data quality level of Q^D. We assume that the values are missing completely at random (MCAR); hence, the incompleteness distributes evenly between the two classes, and the training set contains $Q^D N$ samples of the each group. We annotate the "positive" and "negative" training sets with the missing values by $\{x^1_n\}$ and $\{x^0_n\}$, respectively (in both classes the index n goes between $1..Q^D N$). Under the MCAR assumption, we can use unbiased estimators for the means and the variance:

$$\hat{\mu}_1 = \frac{\sum_{n=1}^{Q^D N} x_n^1}{Q^D * N}, \quad \hat{\mu}_0 = \frac{\sum_{n=1}^{Q^D N} x_n^0}{Q^D * N}$$

$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{n=1}^{Q^D N} \left(x_n^1 - \hat{\mu}_1 \right)^2}{Q^D * N - 1}}, \quad \hat{\sigma}_0 = \sqrt{\frac{\sum_{n=1}^{Q^D N} \left(x_n^0 - \hat{\mu}_0 \right)^2}{Q^D * N - 1}} \tag{11}$$

$$\hat{\sigma} = \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}{2}} = \frac{\sqrt{\sum_{n=1}^{Q^D N} \left(x_n^1 - \hat{\mu}_1 \right)^2} + \sqrt{\sum_{n=1}^{Q^D N} \left(x_n^0 - \hat{\mu}_0 \right)^2}}{2\sqrt{Q^D * N - 1}}$$

As mentioned earlier, if distribution parameters are known, the classification threshold can be calculated by $A=0.5*(\mu_0+\mu_1)$. Here, we need to estimate $\hat{A}$, based on the training set. As the samples in the training set are drawn from Normally-distributed populations, the estimator $\hat{A}$ is also a normally-distributed random variable, for which we can calculate the expected value $E[\hat{A}]$, and the variance $VAR[\hat{A}]$:

$$\hat{A} = \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}, \qquad E\left[\hat{A}\right] = E\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\mu_1 + \mu_0}{2}$$

$$VAR\left[\hat{A}\right] = VAR\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\sigma^2}{2Q^D * N} \tag{12}$$

As discussed in the previous section, the rate of missing values (as reflected by data quality measurement Q^D) may directly affect the classification rule, by increasing uncertainty about best classification threshold. As seen in equation 12 above, missing values that follow the MCAR, do not bias of expected threshold (the expression $E[\hat{A}]$ does not depend on the data quality level Q^D). However, missing values might affect estimation uncertainty and hence, the model quality Q^M. The estimation variance $VAR[\hat{A}]$ and the associated confidence interval (CI), increase with a higher rate of missing values (lower Q^D). As the estimator for the threshold parameter has a Normal distribution, the confidence interval CI_A for the estimator $\hat{A}$, given a desired confidence level ρ, N samples, and a data quality level of Q^D is:

$$CI_A(\rho,N)=\left[\hat{A}-t_{1-\rho/2,2N-2}*\sqrt{\frac{\hat{\sigma}^2}{2Q^D*N}},\quad \hat{A}+t_{1-\rho/2,2N-2}*\sqrt{\frac{\hat{\sigma}^2}{2Q^D*N}}\right] \quad (13)$$

Where,

$\hat{A}$ -	The estimation of the LDA threshold A
ρ -	The target confidence level
N -	The number of samples in the complete training dataset
Q^D -	The data quality level (i.e., the rate of non-missing values)
$t_{1-\rho/2, N}$	The 1-ρ quantile of Student-t distribution with N degrees of freedom

Accordingly, we can calculate the CI-length (and, with equation 2, also the MQ metric) for the LDA threshold A, given a desired confidence level ρ, N samples in the complete dataset, and a DQ level of Q^D:

$$Q_A{}^M\left(\rho,N,Q^D\right)=L_A\left(\rho,N,Q^D\right)=2*t_{1-\rho/2,2N-2}*\sqrt{\frac{\hat{\sigma}^2}{2Q^D*N}}=$$

$$t_{1-\rho/2,2N-2}*\frac{\sqrt{\sum_{n=1}^{Q^DN}\left(x_n^1-\hat{\mu}_1\right)^2}+\sqrt{\sum_{n=1}^{Q^DN}\left(x_n^0-\hat{\mu}_0\right)}}{\sqrt{Q^D*N-1}} \quad (14)$$

Figure 4 shows the model quality (Q^M – the confidence interval length) versus the data quality (QD) for different sample sizes, and with ρ = 0.05. The samples were taken from two normally-distributed populations with μ0=2, μ1 =4 and common σ =3.

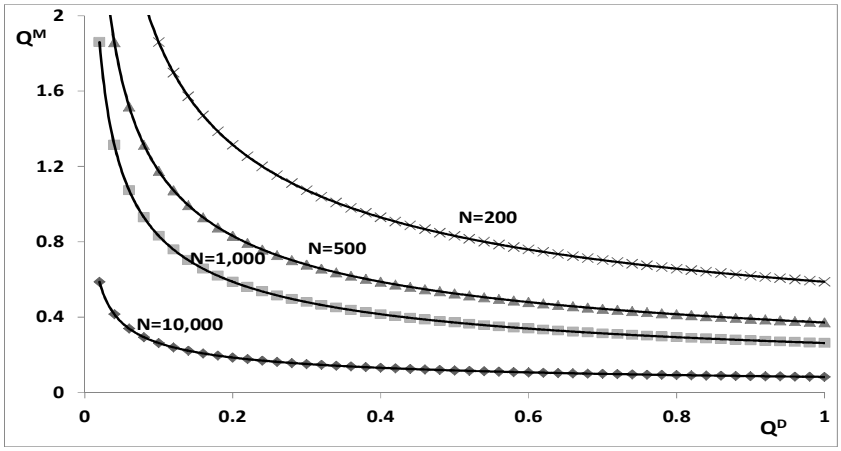

Figure 4: Model Quality (QM) versus Data Quality, with ρ = 0.05

The figure highlights our earlier arguments - model quality is likely to increase (smaller confidence interval) with a higher N, and with a higher DQ level. Notably, with the highest sample-size shown (N=10000), the QM degradation is relatively minor for small QD degradation (QM (QD=1) = 0.08, versus QM (QD=0.6) = 0.1), but becomes more severe as QD reaches low rates (QM (QD=0.1) = 0.26). It can be shown that with a large N, the Student-t distribution can be approximated with a Normal distribution - e.g., with 30 of more degrees of freedom, the error of approximating the probability density function (PDF) of a Student-t distribution with a Normal distribution is less than 0.005. Accordingly, the CI-length will be approximated by $L_A(\rho) = 2* Z_{1-\rho/2} * \hat{\sigma}$.

Decision Quality Metric for the Binary LDA Classifier

After showing the effect of DQ on MQ, we now show the impact of DQ and MQ on the decision quality CQ. In our decision scenario (Table 3), there is no value for correct classification, but some negative cost U for misclassification – hence, we assess decision quality in terms of lowering cost. With known distribution parameters, the lowest-possible cost (Equation 10) was shown to be $C^{C*}= U*\Phi((\mu_1-\mu_0)/2\sigma)$. In this section, we will show that when the parameters have to be estimated from a sample – the decision quality will degrade (i.e., higher negative cost) with a smaller sample size and lower DQ level.

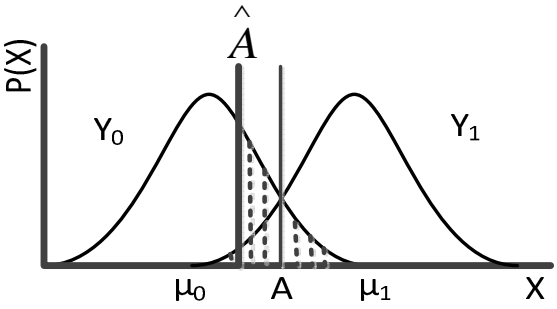

Figure 5: Misclassification Due to Biased Threshold Estimation

Given a certain threshold Â that was estimated from a training set (Equation 14) - misclassification of instance X occurs when it is "positive", but smaller than Â or "negative" but greater than Â. Given a cost parameter of U and an estimated threshold Â, the expected misclassification cost at is:

$$C^C\left(\hat{A}\right)=U*\left(P\left(X<\hat{A}\middle|X\in Y_1\right)+P\left(X>\hat{A}\middle|X\in Y_0\right)\right)=$$

$$U*\left(\Phi\left(\left(\hat{A}-\mu_1\right)\middle/\sigma\right)+1-\Phi\left(\left(\hat{A}-\mu_0\right)\middle/\sigma\right)\right)= \quad (15)$$

$$U*\left(\Phi\left(\left(\hat{A}-\mu_1\right)\middle/\sigma\right)+\Phi\left(\left(\mu_0-\hat{A}\right)\middle/\sigma\right)\right)$$

It can be shown that C^C is minimized when $\hat{A}=A=0.5*(\mu_0+\mu_1)$ (i.e., with a sample size $N\rightarrow\infty$):

$$C^{C^*}=C^C\left(\hat{A}=A=0.5*\left(\mu_0+\mu_1\right)\right)=U*\Phi\left(\left(\mu_1-\mu_0\right)/2\sigma\right) \quad (16)$$

Given a finite sample-size N and a quality level Q^D (i.e., an actual sample size of Q^D*N) – we define the expected classification cost C^c as the mean of $C^c(\hat{A})$ for all possible values of the estimated threshold Â.

$$C^C=E\left[C^C\left(\hat{A}\right)\right]=U*E\left[\Phi\left(\left(\hat{A}-\mu_1\right)\middle/\sigma\right)+\Phi\left(\left(\mu_0-\hat{A}\right)\middle/\sigma\right)\right] \quad (17)$$

The calculation of the mean depends on a certain confidence interval CI – given an actual sample size of Q^D*N, with a confidence rate of ρ (i.e., a likelihood of 1-ρ), the estimated threshold Â will reside within a Δ range around A, where Δ depends on N, Q^D, and ρ.

$$C^C(\rho, N, Q^D) = E\left[C^C\left(\hat{A}\right) \middle| \hat{A} \subset CI \right] =$$

$$U * \frac{\int_{\hat{A} \subset CI} \left(\Phi\left(\left(\hat{A} - \mu_1 \right) \middle/ \sigma \right) + \Phi\left(\left(\mu_0 - \hat{A} \right) \middle/ \sigma \right) \right) d\,\hat{A}}{(1-\rho) * \Delta(\rho, N, Q^D)} = U * \delta(\rho, N, Q^D) \qquad (18)$$

$$where \quad CI = \left[\hat{A} - 0.5 * \Delta(\rho, N, Q^D), \hat{A} + 0.5 * \Delta(\rho, N, Q^D) \right]$$

The expression δ(ρ, N, QD) in Equation 18 reflects the average likelihood that a certain item will be misclassified, given certain values of confidence level ρ, training-set size N, and DQ level QD. It is likely to decrease with a smaller ρ, larger N, and/or larger QD. Figure 6 shows the expected classification cost (CC) versus the data quality (QD) for different sample sizes, with U=1 and ρ=0.05 (the same training sets that were used in Figure 4 - μ0=2, μ1 =4, σ =3).

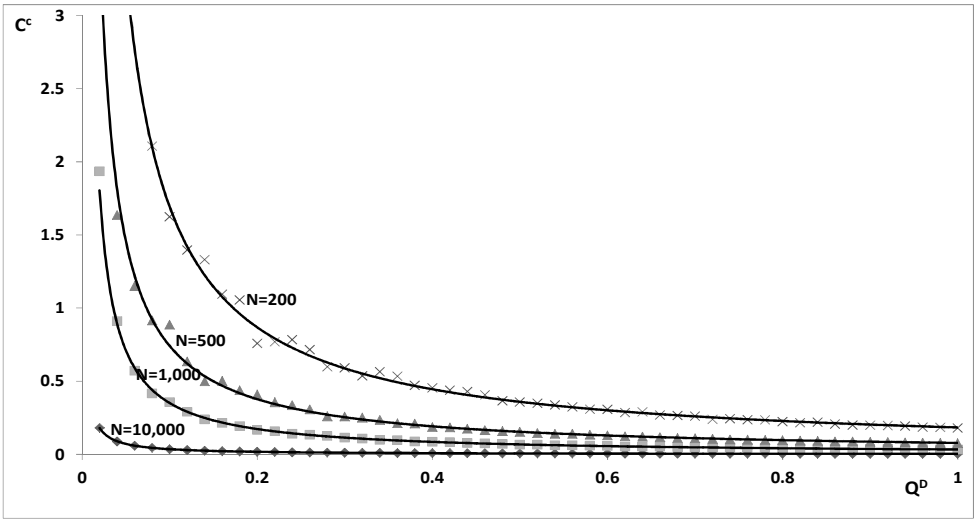

Figure 6: Model Quality (QM) versus Data Quality, with U=1 and ρ = 0.05

The similarity in behavior between Figure 4 and Figure 6 is noticeable – the expected cost is higher with lower sample size, and decreases further as the rate of missing values increases (lower Q^D). With a very large N (here, the maximum take is N=10,000), and with no missing values (Q^D=1), the expected C^C nearly reaches the optimum ($C^{C*} \approx 0.036$). At this large sample size the impact of missing values is relatively minor – there a significant change in C^C only when Q^D goes below 0.1.

Data Quality, Decision Quality and Cost-Benefit Tradeoffs

Assuming that we now have the ability to complete missing values in our training set, at a cost of S units per missing items – would the benefits gained from completing those values justify the associated cost? The answer would be yes – if the reduction in misclassification cost will be higher than the cost of missing-values completion.

Assume that the current quality level is $Q^{D/S}$, and the target quality level is $Q^{D/T}$. If we have N^T items that need to be classified, the classification costs that will be saved by filling in missing values will be

$$\Delta C^C\left(Q^{D/T}\right) = N^T * \left(C^C\left(\rho, N, Q^{D/T}\right) - C^C\left(\rho, N, Q^{D/S}\right)\right) = N^T * U * \left(\delta\left(\rho, N, Q^{D/T}\right) - \delta\left(\rho, N, Q^{D/S}\right)\right) \quad (19)$$

The correction cost ΔC^S of increasing the quality level from $Q^{D/S}$ to a target quality level of $Q^{D/T}$ is:

$$\Delta C^S\left(Q^{D/T}\right) = S * N * \left(Q^{D/T} - Q^{D/S}\right) \quad (20)$$

The net-benefit associating with missing-value competition is given by $B(Q^{D/T}) = \Delta C^S(Q^{D/T}) - \Delta C^S(Q^{D/T})$. We can now frame the question of what quality-level to target as an optimization problem:

Choose $Q^{D/T}$ that maximizes:

$$B\left(Q^{D/T}\right) = N^T * U * \left(\delta\left(\rho, N, Q^{D/T}\right) - \delta\left(\rho, N, Q^{D/S}\right)\right) - S * N * \left(Q^{D/T} - Q^{D/S}\right) \quad (21)$$

S.t., $Q^{D/T} \le Q^{D/T} \le 1$, $B \ge 0$

Where,

B	The net-benefit associated with data quality improvement
$Q^{D/T}$ -	The target data quality level
$Q^{D/S}$ -	The given data quality level
ρ -	The target confidence level
N -	The number of samples in the complete training dataset
N^T -	The number of samples to be classified
$\delta(\rho, N, Q^D)$	The average likelihood of misclassification
U	The expected cost of misclassifying a single item
S	The cost of fixing a single missing value

The objective function formulation in Eq. 21 in not linear and, obviously, does not have a close-form solution; however, the optimal solution can be approximated using a software-based optimization tool. As highlighted by a few studies (e.g., Ballou et al., 1998; Heinrich et al., 2009, Even et al., 2010) – DQ management decisions often involve substantial cost-benefit tradeoffs. The need for cost-benefit assessment is also reflected in the analysis done in this study – but with some separation between the datasets on which we act. The data correction cost is associated with the training set, used for building the model. On the other hand, the reduction in misclassification cost is associated with data items that are not part of the training set, but have to be classified according to the model developed.

Discussion - Limitations and Future Extensions

The general methodology described earlier suggests that DQ may affect MQ, and hence CQ behavior. This section developed this argument further by demonstrating an analytical methodology that shows the explicit link between the three levels. This section introduced a more detailed development of that concept for binary LDA classifiers – a relatively simple, yet useful classifier. The development showed explicit and quantifiable links between the missing-value rate (as reflected by the DQ measure Q^D), the model quality (in terms of minimizing the confidence-interval length), and the decision quality (in terms of minimizing misclassification costs). As shown in Equation 21, the mapping between the data quality level and the expected misclassification cost can be used for developing analytical tools that permit cost-benefit assessments. Based on the results of such assessments – the target quality level can be set, such that the margin between the classification-cost saved and the correction cost will be maximized.

To highlight the key concepts and arguments – the analytical development in this section was done under some simplifying and restrictive assumptions. Those assumptions should be relaxed in future extensions to this study, as summarized in Table 4.

Issue	Assumption Made	Future Extensions
Dimensions	• Scalar ("1-dimensional") input	• Multidimensional input vector
Classes	• Two	• Any $K \geq 2$
Symmetry	• Class distributions with different means, but identical STDEV • Same a-priory probability • Same number of samples per class • Same misclassification cost for "false positive" and 'false negative"	• Asymmetry between classes in terms of standard deviations, a-priory probabilities, sample size, and misclassification costs
Distribution	• Normal	• Other distributions, not necessarily symmetric
Classifier Type	• Linear, based on a separating hyper-plane	• Non-linear, based on more complex separation rules - e.g., Quadratic Discriminant Analysis (Duda and Hart, 2001)
Missing-Values Pattern	• Missing completely at random (MCAR)	• Patters with certain non-random associations between missing values (e.g., MAR – Missing at Random; NMAR – Not missing at Random (Little, 1987))
DQ Criterion	• Missing-value defects	• Other DQ defect types – e.g., inaccurate, invalid, and/or outdated data items
MQ Criterion	• Confidence interval, calculated per parameter	• Other criteria that consider the entire model
CQ Criterion	• Minimizing classification cost	• Maximizing accuracy, precision, sensitivity, and/or specificity • Maximizing classification value
Decision Scenario	• Classification, based on a discrete set of classes	• Optimization – setting the optimal value within a continuous value range

Table 4: Assumptions and Future Extensions

CONCLUSIONS

The negative impact of DQ defects on decision making has been broadly acknowledged in research and in practice. This study suggests that a possible way to understanding and quantifying this impact is by looking into the mediating role played by decision-support models. Such models are often estimated from training datasets – and when such a training dataset suffers from DQ defects, the model and the decisions that it supports are likely to be biased. This claim makes intuitive sense – however, not much was done to support it analytically. This study takes a step in that direction by offering an analytical framework that links the three levels of quality assessment - data quality, model quality, and decision quality. The analytical development demonstrated in this study is relatively simple – and its aim was to highlight and demonstrate the key concepts. As this study is still progressing – our goal is to examine comprehensive and complex decision scenarios, in which some of the assumptions made will be relaxed.

REFERENCES

[1] Askira-Gelman, I. GIGO or not GIGO: The Accuracy of Multi-Criteria Satisficing Decisions, The ACM Journal of Data and Information Quality, 3(2), Article 9, 2011, pp. 1-27

[2] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.

[3] Blake, R., and Mangiameli, P., The Effects and Interactions of Data Quality and Problem Complexity on Classification. ACM Journal of Data and Information Quality, 2 (2), Article 8, 2011, pp. 1-28

[4] Davenport, T.H. Competing on Analytics. *Harvard Business Review*, 84(11), 2006, pp. 99-107

[5] Duda, P. E. and Hart, D.G. S., *Pattern Classification*, Wiley & Sons, 2001.

[6] Even, A., and Shankaranarayanan, G. Utility-Driven Assessment of Data Quality, SIGMIS Database, 38(2), 2007, pp. 76-93

[7] Even, A., Shankaranarayanan, G., and Berger, P.D. Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting, *Decision Support Systems,* 50(1), 2010, pp. 152-163

[8] Han, J. and Kamber, M. *Data Mining – Concepts and Techniques*, Elsevier, 2006

[9] Heinrich, B., Kaiser, M. and Klier, M. A Procedure to Develop Metrics For Currency And Its Application In CRM, *The ACM Journal of Data and Information Quality*, 1(1), 2009 pp. 5-28

[10] Little R.J.A. *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987

[11] Liu, S., Duffy, A., Whitfield, R., and Boyle, I. Integration of Decision Support Systems to Improve Decision Support Performance. *Knowledge and Information Systems*. 22(3), 2010, pp. 261-286

[12] McLachlan J. G. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 1992

[13] Parssian, A., Sarkar, S., and Varghese, S.J. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science, 50(7), 2004, pp. 967-982

[14] Pipino, L.L., Lee, Y.W., and Wang, R.Y. Data Quality Assessment, *Communications of the ACM*, 45(4), 2002, pp. 211-218.

[15] Redman, T.C. *Data Quality for the Information Age*, Artech House, 1996

[16] Sessions, V., and Valtorta, M. Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm, Journal of Data and Information Quality, 1(3), Article 14, 2009, pp. 1-34

[17] Shankaranarayanan G, and Kay Cai Y. Supporting Data Quality Management in Decision-Making. *Decision Support Systems*, 42(1), 2006, pp. 302-317

[18] Shim, J.P., Warkentin, W., Courtney, J.F., Power, D. J., Sharda, E., Carlsson, C. Past, Present, and Future of Decision Support Technology, *Decision Support Systems*, 33, 2002, pp. 111-126